Structuring Clinical Practice Guidelines in a Relational Database Model for Decision Support on the Internet

David F. Lobach, MD, PhD, MS; Cynthia S. Gadd, PhD, MBA; and Joseph W. Hales, PhD
Division of Medical Informatics, Department of Community and Family Medicine
Duke University Medical Center, Durham, North Carolina

The rapid proliferation of clinical practice guidelines (CPGs) has made computerization increasingly useful to clinicians. Computerization, however, requires transformation of the content and logic of each guideline into a computer-accessible form. In this project, we sought to use a relational database to construct a generalized guideline knowledge base for use with Internet-based decision support applications. We hypothesized that knowledge representation schemes could be developed to capture guideline content and logic within the constraints of a relational database model. In this paper we describe a database schema based on a relational model for computerizing CPGs using a hybrid of structured and procedural knowledge representation schemes. We developed and refined this model in the context of five diverse CPGs and found it accommodated all necessary representational requirements.

INTRODUCTION

Clinical practice guidelines (CPGs) are playing an increasingly important role in standardizing health care [1]. It has become nearly impossible for a clinician to integrate all 1600 of the available CPGs into his/her practice using traditional paper-based systems. Computerization of CPGs has been shown to improve compliance with guideline recommendations and to improve outcomes [2,3]. In order to computerize guidelines, however, guideline content and logic must be represented in a computer-accessible form. Consequently, there is a need to develop general knowledge representation schemes for CPGs in the context of existing database models.

Several approaches of variable complexity have been employed to represent CPG knowledge electronically. Few of these knowledge representation efforts, however, have been used to implement CPGs in clinical practice [4-6]. Previous structured approaches have not used a relational database model [6] or have lacked expressivity for representing temporal and complex knowledge [5]. Procedural knowledge representation approaches have tended to be system-specific [7] or inordinately complex because they have not been explicitly designed for CPG [4]. The Arden Syntax approach, a hybrid of procedural and structured representation schemes with a single rule-single decision modular design, can lead to unexplained interactions between modules because it does not model related decisions well [8]. Arden Syntax requires knowledge to be structured in a complex, non-intuitive format in modules that often are redundant [4,8].

We hypothesized that we could develop a knowledge representation scheme for CPGs that would conform to a relational database model without compromising expressivity or completeness. In this paper, we present a relational database model for CPGs which uses a hybrid of structured and procedural knowledge representation formalisms to represent guideline content and logic. This empirically developed model provides the knowledge base for Siegfried (System for Interactive Electronic Guidelines with Feedback and Resources for Instructional and Educational Development), a research project using the Internet to present interactive CPGs that are customized to an individual patient and available at the point of care.

METHODS

We selected a relational database format to store our CPG knowledge because it is compatible with Internet-based applications; it is a familiar model that has evolved into an industry standard; it is supported by multiple database management tools; it is sharable through applications using Structured Query Language (SQL), and it is easily explained using a tabular representation. To illustrate the development of our knowledge representation scheme and our relational database model, a CPG on low back problems is used as an example throughout this paper [9].

Model Development

We extracted content and logic from five CPGs (Table 1) to drive empirical database model development and knowledge representation. Initially, we created a hypothetical entity-relationship (ER) diagram for CPGs and then reshaped this model using content and logic from the low back problems and high cholesterol management CPGs. The remaining three CPGs were later used to refine the comprehensiveness and expressivity of the model. To prepare a CPG for knowledge extraction, we converted each guideline into a unidirectional graph (see example in Figure 1). Each node in the graph corresponds to a decision point or a recommendation. In order to traverse the graph in a breadth-first search, we defined levels within each graph which specified a group of nodes that must be resolved before proceeding deeper into the algorithm.

Extracting Guideline Content. To represent the *content* we extracted from CPGs in a relational

1091—8280/97/$5.00

Table 1. Clinical Practice Guidelines Selected for Knowledge Representation

Guideline Focus	Source	Type of Care	Attributes for Selection
Acute Low Back Problems	National Government Agency (AHCPR)	Acute Disease Management	Well described algorithms with extensive branching based on real time data input (similar to a care map)
Diabetes Mellitus	Private Specialty Society (ADA)	Chronic Disease Management	Text based description of optimal care requiring translation to algorithms for implementation
Cholesterol Management	Health Maintenance Organization (NYL-Care)	Focused Therapy	Text-based care standards defined by insurers (based on NCEP screening recommendations)
Smoking Cessation	National Government Agency (AHCPR)	Behavioral Modification	Predominantly descriptive, educational content with emphasis on educating providers
Preventive Services	National Government Task Force (USPSTF)	Preventive Health	Mixture of static (not requiring real time input) and active (requiring real time input) care recommendations

database format, we selected a structured knowledge representation scheme implemented through frames. Each row in the relational tables we created for CPG content defined a frame for capturing structured knowledge. To extract content information we wrote a logic statement to correspond with each decision node (Figure 1, Content). From these statements we identified the critical concepts, which we termed *elements*, that were needed to express guideline content, e.g., node 1: "age" and "activity intolerance due to...." We also identified critical values, if any, that were associated with each element in the context of the guideline, e.g., "age ≥ 18" to denote "adult." The guideline elements served to define a set of *primary data parameters* from which the elements could be derived, e.g., "age" and "activity intolerance." Primary data parameters were functionally defined as context-independent clinical data that would be routinely collected in a comprehensive computer-based patient record (CPR). A primary data parameter was placed in context to become an element through any or all of the following approaches: addition of temporal information (e.g., "duration < 3 months"), association with a modifier (e.g., "due to low back problems"), abstraction to a higher level concept, or transformation with a mathematical or logical formula. Temporal and non-temporal transformations were represented separately in the model because they had different attributes. Content extraction also led to a representation for the source of each primary data parameter including upper and lower normal ranges for specific continuous threshold values explicitly used in the CPG logic. Content extraction and representation were completed by assigning elements and primary

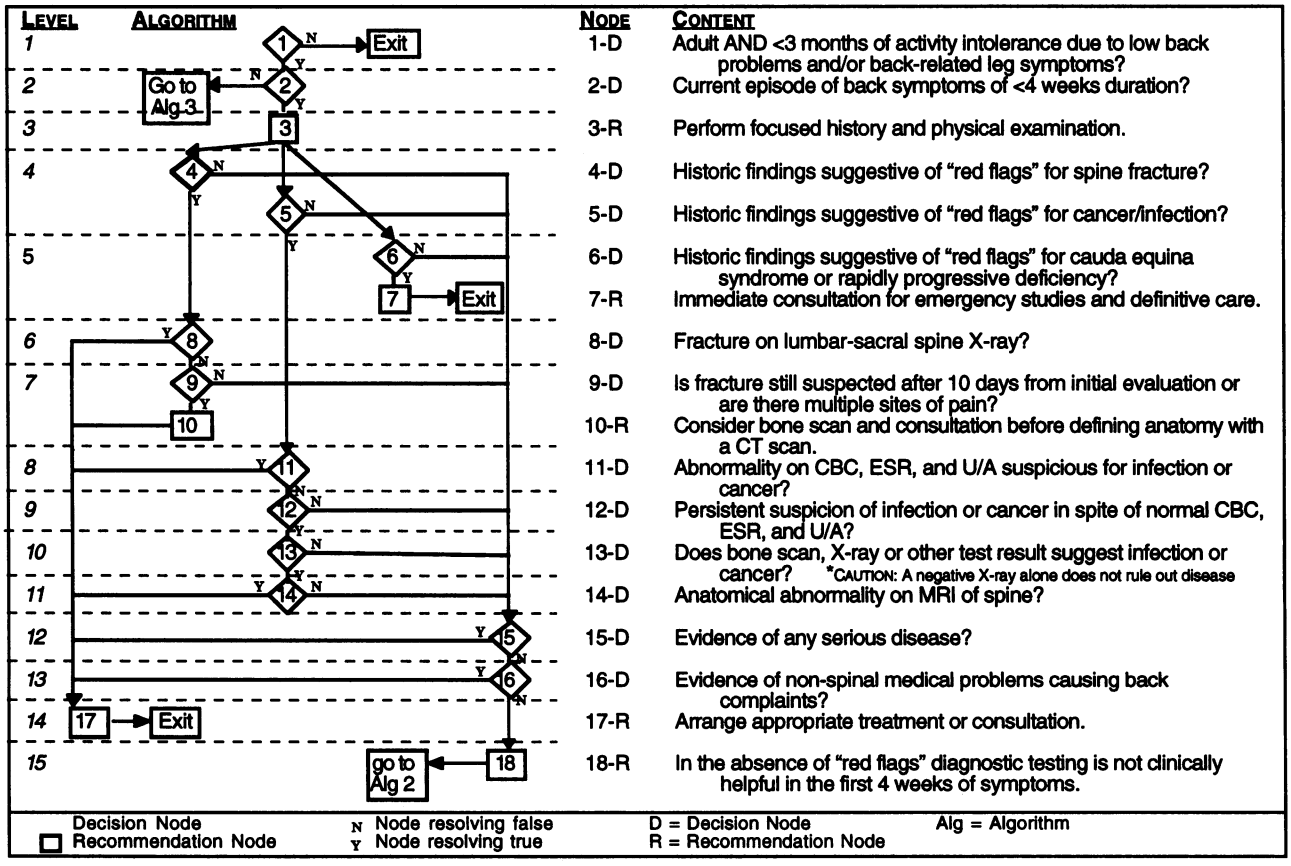

Figure 1. Unidirectional graph derived from a clinical practice guideline for low back problems.

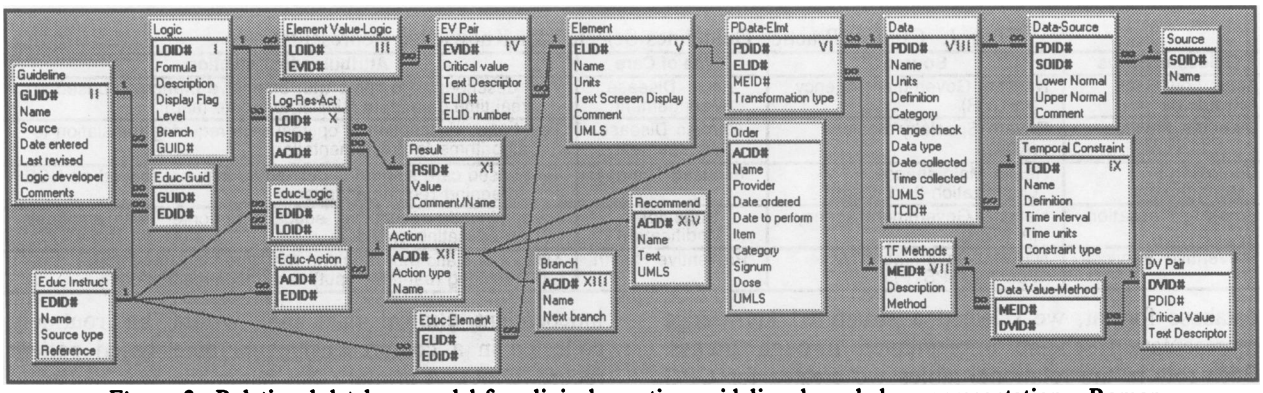

Figure 2. Relational database model for clinical practice guideline knowledge representation. Roman numerals in the upper right corners of selected frames correspond to instantiated tables in Figure 3.

data parameters to a concept in the Unified Medical Language System (UMLS) whenever possible. If a UMLS code was not available, we assigned an internal code derived from the name of the relational key and the unique key identifier.

Extracting Guideline Logic. The logic extraction process resulted in the selection of a procedural knowledge representation scheme for *logic-related* guideline knowledge. We converted each logic statement into an internally defined formalism in which an operator (=,<,>,≠) was used to define the relationship between a guideline element and its critical value, and element/critical value pairs were grouped with the appropriate Boolean operators (e.g., ((≥ elmt-value pair ID# {age, 18}) AND (= elmt-value pair ID# {activity intolerance due to ... of 3 months duration, present})). Since some logic statements required identification of the number of element/critical value pairs that resolved either true or false, we instituted a functional formalism in which each logic statement returned a count of the number of element-value pairs that were either true or false depending on the specification in the function, e.g., the traditional "AND" operator was represented as COUNT_FALSE ((x) (y) (z)) and resolved true if it returned a zero value. In order to maintain context for the logic, we associated branch and level information with each logic statement, e.g., node 1:branch 1:level 1. The next step in logic representation entailed identifying all possible results from processing the logic statement, and the actions paired with each of these possible results. These actions were then classified as recommendations, orders or branches. Branch actions served to direct traversal of the guideline algorithm by specifying which branch should be accessed next. The final phase of the logic representation process involved assigning UMLS or internally derived codes to selected actions.

The last component of knowledge extraction from a CPG entailed creation of a list of instructional or educational resources that supported the guideline. These resources were then paired with the appropriate logic statement, element, action or guideline.

RESULTS

The ER diagram corresponding to our generic relational database model for CPGs is shown in Figure 2. This model has 24 relational tables and 85 unique attributes. Each row in the logic table corresponds to a decision node from the CPG graph (Figure 1). A description of the representation accommodated by selected tables and a sample data entry for these tables is depicted in Figure 3. The logic table is listed first since it dictates the content (i.e., elements, critical values, primary data, transformations, etc.), the actions, and the educational resources associated with each guideline. An understanding of the sequential progression through these tables can be enhanced by matching each table in Figure 3 with the corresponding table in the ER diagram (Figure 2). Traversing the relationships in the ER diagram serves to organize the tables and data in Figure 3 into the relational paradigm.

During the course of model development and testing, we encountered several challenges which served to shape the model. The first challenge was representing time-oriented concepts. We developed the temporal constraint table to accommodate temporal context, ordering and duration through the "definition," "time interval," "time units" and "constraint type" fields [4]. The temporal duration of an event such as a symptom that is present for <3 months can be specified by entering the pivotal interval of occurrence ("3"), the time units ("months"), a constraint type designating when it occurred ("been present") and a definition of how the interval should be interpreted ("no longer than past"). Operations and comparisons using temporal data for determining ordering and context are handled by combining elements with duration assigned to each, e.g., back problems for 3 months. A second challenge was creating a logic formalism that determined not only if a complex logic statement resolved true or false, but also how many components were either true or false. We implemented a LISP-like function which returned a numeric value. This value could then be compared to a threshold value in the result table to resolve the logic statement true or false (see logic and

i. LOGIC *(formulas for each decision node and description of logic statement)*

LOID#	Formula (operator EVID#)	Description	Branch	Level	GUID#
0002	(COUNT_FALSE (≥0001)(=0002))	age≥18 and activity intol < 3 months	1	1	0001
0003	(COUNT_TRUE (≤0003)	back symptoms < 4 weeks duration	2	2	0001

ii. GUIDELINE *(source and other characteristics associated with a clinical practice guideline)*

GUID#	Name	Source
0001	acute low back problems	Agency for Health Care Policy and Research

iii. ELEMENT-VALUE-LOGIC *(junction table listing EV pairs used as parameters in logic statement)*

LOID#	EVID#
0002	0001
0002	0002

iv. EV PAIR *(association of an element with a critical value)*

EVID#	Critical Value	Text Descriptor	Element Name	ELID#
0001	18	adult age	age	0001
0002	Y	present	activity intolerance < 5 months duration	0003

v. ELEMENT *(concepts placed in context for use in the guideline logic)*

ELID#	Name	Unit	Screen Text Display	UMLS
0001	age	years	age in years	ELID0001
0003	activity intolerance, <3 months	--	intolerance to activity due to low back problems and/or back related leg symptoms	ELID0003

vi. PDATA-ELMT *(junction table for many-to-many relationship and association with transformation method)*

PDID#	ELID#	MEID#	Transformation Type
0001	0001	--	--
0002	0003	0001	modifier

vii. TF METHODS *(procedures implemented to change primary data parameters into elements)*

MEID#	Method
0001	ADD MODIFIER: due to low back problems and/or back related symptoms

viii. DATA *(listing of primary data parameters and their attributes)*

PDID#	Name	Units	Definition	UMLS	TCID#
0001	age	years	age of individual	C0001779	0001
0002	activity intolerance	--	intolerance to activity	C0150008	0010

ix. TEMPORAL CONSTRAINT *(list of methods applied to primary data parameters to represent time)*

TCID#	Name	Definition	Time Interval	Time Units	Constraint Type
0001	current	at this time	--	--	is/are present
0010	<3 months	no longer than past	3	months	been present

x. LOG-RES-ACT *(junction table for tertiary many-to-many relationship)*

LOID#	RSID#	ACID#
0002	0001	0002
0002	0002	0003
0002	0002	0004

xi. RESULT *(threshold values for interpreting logic statement true and false counting functions)*

RSID#	Value	Comment/Name
0001	=0	--
0002	≥1	--

xii. ACTION *(Superclass listing of actions that follow from resolution of logic statements)*

ACID#	Action Type	Name
0002	branch	node 1
0003	recommendation	--
0004	branch	node 1

xiii. BRANCH *(list of actions that dictate how an algorithm is traversed)*

ACID#	Name	Next Branch
0002	node 1	2
0004	fails criteria	EXIT

xiv. RECOMMENDATION *(list of actions that stipulate recommendations to be presented)*

ACID#	Name	Text	UMLS
0003	activity intolerance >3 months	Patient does not meet criteria for this low back problems guideline	ACID0003

Figure 3. Relational model content description and sample knowledge entry for selected tables from a clinical practice guideline on low back problems. Format of figure entries: *line 1*--TABLE NAME *(representational function)*, *line 2*--Selected field names, *line 3* --Sample data representing a specific logic statement (*Adult patient and < 3 months of activity intolerance due to low back problems; source: Figure 1, level 1, node 1*). Four digit numbers are internal indices.

and result tables in Figure 3). A third challenge was to create a model that not only supported content and logic to implement guidelines, but also contained information to display the guideline interactively to a user. Text fields were incorporated into tables to provide a complete explanation of the guideline in structured language. The synthesis of the text fields into comprehensible English is handled by the guideline inference and display engine. A sample display for the logic statement in Figure 3 is:

Select true statements:
- ❑ Age in years ≥ 18 is/are present
- ❑ Activity intolerance due to back problems and/or back related symptoms has been present no longer than past 3 months

A fourth challenge was the development of a representation system that linked concepts adapted to the context of a CPG with primary data parameters that would be collected in a CPR. We achieved this representation by creating a table for contextualized elements that were directly used in the guideline and a table that linked these elements with primary data parameters and the methods required to transform these parameters into elements. Temporal representation was described above. Non-temporal transformations included descriptive modifiers, e.g., "due to back problems;" algebraic formulas; logic statements; and abstraction or synthesis of concepts, e.g., determining best estimated gestational age from last menstrual period, fundal height and fetal ultrasound data. Specifications for each type of transformation were procedurally defined and resolved by the guideline inference engine. A final challenge was finding a controlled vocabulary to represent guideline concepts. A controlled vocabulary was needed to avoid duplication of concepts in the model and to facilitate mapping of guideline terms to data to be imported from a CPR in a future project. UMLS was selected because it is one of the most comprehensive medical concept representation schemes available and because concepts extracted directly from CPG are incorporated into its knowledge base [10].

DISCUSSION

In this paper we describe a relational database model that captures knowledge from CPGs using a structured knowledge representation scheme for guideline content and a procedural representation scheme for guideline logic. The model and representation schemes were initially derived using two CPGs and, subsequently, refined with 3 additional CPGs. Informally, the model

has been tested as a suitable representational format against several other CPG as well as care maps and patient questionnaires. Through both its refinement and informal testing, the model has successfully accommodated the needed representations.

This database model is noteworthy because it fully supports knowledge representation schemes for CPG content and logic in a relational format. The relational format facilitates integration with Internet-based applications and database tools which will be part of the Siegfried project. It also uses a familiar, intuitive database paradigm that is amenable to knowledge entry without programming [11]. For easier maintenance and modification, knowledge entered in the relational database remains readable in its frame-based representation. Additionally, to facilitate scalability and model changes over time and to permit addition of new guidelines, content knowledge and control knowledge are partially separated and the knowledge base is external to the inference and display engine [4].

Our concept-oriented, frame-based design allows for internal sharing of all data types except for logic statements which are specifically linked to a guideline. Shared concepts will make data modifications easier since concepts are not duplicated at multiple places throughout the knowledge base. Because guidelines are linked to logic, and logic is linked to elements, and elements are related to primary data parameters, the primary data parameters required for a guideline can be easily identified. These primary data parameters can then be requested from a CPR when a guideline is activated. Other unusual features of this model include the incorporation of display related information and links to educational resources.

Our model is limited by the fact that it has been tested formally against only five CPGs. We have sought to maximize the robustness of this testing by selecting diverse CPGs with diversified representational requirements. Further testing is warranted and could lead to minor changes in the model which the relational format should readily accommodate. The model is also limited because it provides no internal mechanism to deal with the ambiguities which are characteristic of many guidelines. To handle these ambiguities, we have elected to present ambiguous guideline components interactively to the user to allow him or her to review and resolve the ambiguity.

In summary, we have demonstrated that the complete content and logic from five CPGs can be structured in a relational database model using a hybrid of structured and procedural knowledge representation. The general applicability of this representation and model will accommodate expansion to include multiple guidelines. We are currently developing a generic, Web-based inference and display engine that will present guidelines from the knowledge base interactively to users at the point of care.

Acknowledgments

This work was funded in part by R01 HS09436-01 from the Agency for Health Care Policy and Research.

References

1. Audet AM, Greenfield S, Field M. Medical practice guidelines: current activities and future directions. *Ann Intern Med.* 1990;113:709-714.
2. Lobach DF, Hammond WE. Computerized decision support based on a clinical practice guideline improves compliance with care standards. *Amer J Med.* 1997;102: 89-98.
3. Tierney WM, Overhage JM, Takesue BY, et al. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. *JAMIA.* 1995;2:316-22.
4. Musen MA, Tu SW, Das AK, Shahar Y. EON: A component-based approach to automation of protocol-directed therapy. *JAMIA.* 1996;3:367-88.
5. Kuperman GJ, Teich JM, Bates DW, McLatchey J, Hoff TG. Representing hospital events as complex conditionals. In: Gardner RE, ed. *19th SCAMC.* Philadelphia: Hanley & Belfus, Inc. 1995; 137-41.
6. Barnes M, Barnett GO. An architecture for a distributed guideline server. In: Gardner RE, ed. *19th SCAMC.* Philadelphia: Hanley & Belfus, Inc. 1995;233-7.
7. Overhage JM, Mamlin B, Warvel J, Warvel J, Tierney W, McDonald CJ. A tool for provider interaction during patient care: G-CARE. In: Gardner RE, ed. *19th SCAMC.* Philadelphia: Hanley & Belfus, Inc. 1995;178-82.
8. Sherman EH, Hripcsak G, Starren J, Jenders RA, Clayton P. Using intermediate states to improve the ability of the Arden Syntax to implement care plans and reuse knowledge. In: Gardner RE, ed. *19th SCAMC.* Philadelphia: Hanley & Belfus, Inc. 1995;238-42.
9. U.S. Department of Health and Human Services. *Clinical Practice Guideline Number 14: Acute Low Back Problems in Adults: Assessment and Treatment.* Rockville, MD: Agency for Health Care Policy and Research, 1994.
10. Humphreys BL, Hole WT, McCray AT, Fitzmaurice JM. Planned NLM/AHCPR large-scale vocabulary test: using UMLS technology to determine the extent to which controlled vocabularies cover terminology needed for health care and public health. *JAMIA.* 1996;3:281-287.
11. Hales JW, Gadd CS, Lobach DF. Development and use of a Guideline Entry Wizard to convert text clinical practice guidelines to a relational format. In Masys DS. *Proceedings: 1997 AMIA Fall Symposium.* 1997; in press.